1 Million Pedestrian Point Cloud Scenes

Michael Saxon⋆, John Kevin Cava⋆, Todd Houghton⋆, Calvin Norman⋆,
Hongbin Yu

Arizona State University

Abstract. We are interested in modeling real-world scenes as 3D point cloud sequences through self-supervised generative representation learning. Successful demonstrations of such learning in other modalities, including vision, speech, and language, have been driven by large, diverse datasets. At present, most readily available point cloud sequence datasets are comprised of vehicular LiDAR data. While these data are valuable we are interested in analyzing other kinds of scenes, particularly ones with a stationary frame of reference. To this end we have produced the ASU Pedestrian LiDAR Scenes (APLS) dataset, containing over 1 million stationary LiDAR pedestrian crowd frames, in 10 different locations around the campus of Arizona State University. To our knowledge this is the first stationary pedestrian LiDAR scene dataset of this size, containing over 30 hours of data collected from a Velodyne HDL-32E LiDAR sensor, totalling over 500 GB of data.

Keywords: LiDAR, dataset, point cloud sequence, representation learning, pedestrian crowd

1 Introduction

The intuition that some understanding undergirds the ability to create has driven work on generative representation learning across a variety of domains. Generative adversarial networks and variational autoencoders have pushed the envelope in generating original, photorealistic images, from distributions of similar images or text [1]. Language model-based feature extractors and transformers pretrained on language-modeling tasks [2] have been successfully deployed in both text generation and text labeling tasks. In these cases, the underlying representation learning has been driven by the abundance of usable data. To learn sophisticated distributions data is often collected en masse from the internet, cleaned, and labeled; such was the case with ImageNet, English Wikipedia, and Common Crawl[1].

We are interested in modeling active 3D scenes in time with representation learning; however limited data for such a task is available. LiDAR point cloud data is well-suited for this purpose. Unfortunately, LiDAR datasets are relatively rare and small at present, and do not contain diverse scenes.

⋆ Equal contribution

[1] https://commoncrawl.org/

Most available LiDAR data are from autonomous vehicle datasets, collected from sensors atop moving roadway vehicles. Point cloud data acquired from a moving vehicle is not suitable for learning representations of evolving scenes, and represent only a small subset of the kinds of scenes present in urban environments. Such scenes contain road-specialized features like pavement, intersections, traffic signs, medians, and sidewalks. Results from other modalities suggest that a diversity of samples is critical for training robust representations; thus point cloud sequences depicting other sorts of scenes are desirable. We are particularly interested in modeling pedestrian scenes, as they have been studied in the vision modality already as a good source of information on the social dynamics of people moving through crowded scenes [3].

2 Related Work

For a time available LiDAR datasets were mostly small. The following autonomous vehicle datasets contained moving reference frame Velodyne LiDAR sequences, and were released before 2019. Polyterrasse and Tannenstrasse, contained 900 and 500 Velodyne HDL 64E point clouds respectively [4]. KITTI, contains 8,039 moving vehicle frames [5]. Ford Campus Vision and LiDAR contains on the order of 5000 LiDAR frames, although the authors do not state exactly how many [6]. L-CAS contains 49 minutes of data [7]. The ApolloScape dataset contains 165,949 semantically-labeled paired LiDAR-video frames [8].

Beginning in 2019, several larger-scale vehicular LiDAR datasets were released. The Waymo Open Dataset [9] contains 200,000 10Hz frames in 1,950 20s segments. Each frame contains mid- and short-range LiDAR and 5 front and side camera images. Additionally, they provide 12.6M bounding boxes and 11.8M semantically-labeled frames. A2D2 [10] contains 390,000 unlabeled vehicular LiDAR frames, 40,000 that are semantically labeled, and 12,000 with 3D object bounding boxes. nuScenes [11] contains 1000 20 second scenes for 390,000 total LiDAR sweeps. 23 object classes are provided with 3D bounding boxes at 2Hz over the dataset, for a total of 1.4M bounding boxes. The Lyft Level 5 dataset [12] contains over 1000 hours of Velodyne sweeps with labeled traffic agent motion.

Consequently, work on generative point cloud representation learning [13] and autoencoder modeling [14] for static objects, rather than evolving 3D scenes, has shown promise. One existing real-world point cloud dataset that is well-suited for these models is the Stanford track collection, which contains 14,000 labelled object tracks, including pedestrians, bicyclists, etc, extracted from larger Velodyne sweeps [15]. End-to-end neural object detection using hierarchical representations has been demonstrated on KITTI data [16].

The existing data resources are rich and well-suited for generative scene representation learning–but only on the restrictive domain of vehicle-mounted scanners. For the aforementioned reasons, we are particularly interested in crowded pedestrian scene data. One existing dataset that provides this is the NCLT Dataset, contains 34.9 hours of data from a robot-mounted LiDAR moving

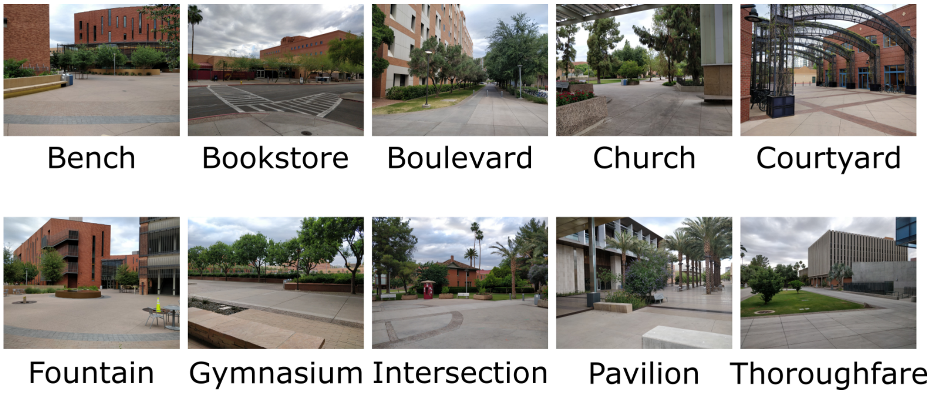

Fig. 1. The ten APLS collection run locations.

around a college campus [17]. We want to diversify the pool of existing point cloud scene data by providing a large resource of stationary reference frame, non-vehicular, non-road data. Our dataset, when considered alongside NCLT, effectively doubles the amount of available college campus pedestrian scene data available.

3 Dataset

The ASU Pedestrian LiDAR Scenes dataset[2] (APLS) dataset consists of 20 collection runs portraying urban pedestrian traffic scenes. The collection runs were completed in 10 different outdoor urban locations at Arizona State University in Tempe, Arizona, as depicted in Figure 1. At each location, 1–4 collection runs were completed, wherein a Velodyne HDL-32E LiDAR unit captured 10Hz sequences of 32 vertical channel 360 degree point clouds, recording the surrounding environment. The resulting data set contains 1,104,314 point cloud frames, constituting 30.68 hours of real-time data. We selected the physical location and time of each collection run to maximize the amount of visible pedestrian traffic. Locations containing numerous motor vehicles were avoided to reduce the amount of spurious objects moving between frames.

The APLS dataset is large, containing over 500 gigabytes of uncompressed raw packet capture `.pcap` data. Because these packets contain spurious data, and must be read sequentially, we break them into "Distance-only Velodyne" DOVe files containing only the distance and position-related binary data, a 50% size reduction. To enable easy use of this dataset we provide Python DataLoader code as well as a parser to convert other Velodyne LiDAR `pcap` files into this format, allowing future LiDAR data to be used with ours.

[2] Dataset available at `https://asulidarset.github.io`

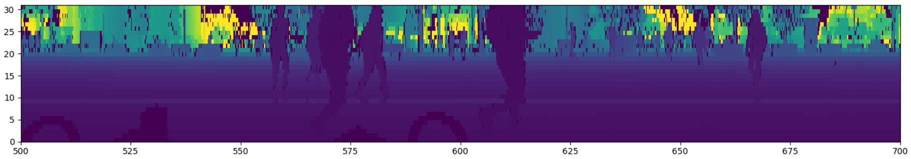

Fig. 2. "Depth camera" rendering of a sample frame of pedestrians at *intersection*.

Table 1. Statistics for each collection location; total duration in HH:MM format.

Location	# Runs	Traffic Level	# Frames	Duration
Bookstore	2	High	112,069	3:06
Church	2	Moderate	116,802	3:14
Courtyard	1	Sparse	92,785	2:34
Thoroughfare	1	Moderate	53,919	1:29
Pavilion	2	Moderate	109,689	3:03
Gymnasium	3	High	107,645	2:58
Intersection	2	Moderate	109,843	3:02
Boulevard	1	Low	55,284	1:32
Fountain	2	High, Moderate	76,825	2:08
Bench	4	Moderate	269,453	7:29
Total	20	—	1,104,314	30:41

4 Future Work

4.1 Refinements to APLS

In its current form, APLS is unlabeled, primarily in order to facilitate work on unspervised learning of scene dynamics in a more pedestrian setting that isn't constainred to motor traffic. However, the addition of labeled data would greatly increase APLS as a resource in terms of possible object detection, object segmentation, object tracking, etc. in data that juxtapose many dynamic entities with a static background.

As such, future work for APLS will definitely try to include bounding box labels, and trajectory paths for said bounding box objects. Moreover, we would want to construct a simpler semnatic label scheme to label the dynamic objects in the scene, especially one that in tuned with a campus environment e.g bikes, pedestrians, skateboards, and more recently motorized scooters.

4.2 Challenges to Unsupervised

Currently there are no baselines that were done for this dataset, as there is currently challenges in determining what would be constitute an effective metric for an unsupervised spatial-temporal scene understanding task.

One challenge is how to we want to frame the prediction, as 2D or 3D. There have been various previous work that do spatial-temporal reconstruction and prediction with video by utilizing CNN-LSTMs. However, LiDAR data are sparse, and trying to reconstruct and predict future frames from sparse data, may be difficult.

Conversely, by framing reconstruction and prediction as a generative point cloud data, this leads to very new work, as we can tell. Previous work have only delt with reconstruction and representations of 3d points of objects, or inside environments. The closest work would be PointRNN [18], which combines PointNet with an RNN framework in order to generate future point cloud frames. However, this work has been done in data such as nuScenes, that leverage the dynamic agent of the car, and also the generative prediction is limited to a distance of 5m around the LiDAR.

5 Conclusion

We argue that by training from a large collection of sequential, unlabeled, LiDAR scenes, learned representations can be leveraged for future prediction problems (i.e pedestrian counting, classification, segmentation, and trajectory tracking).

In this work we presented the ASU Pedestrian LiDAR Scenes dataset, consisting of over 30 hours of point cloud sequence data of pedestrian scenes on a college campus. To our knowledge, this one of the largest publicly available LiDAR point cloud sequence datasets, and the largest for stationary pedestrian scenes. We hope that future work will leverage this resource to realize self-supervised 3D scene representation learning.

References

1. Wu, X., Xu, K., Hall, P.: A survey of image synthesis and editing with generative adversarial networks. Tsinghua Science and Technology **22**(6) (2017) 660–674
2. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
3. Alahi, A., Goel, K., Ramanathan, V., Robicquet, A., Fei-Fei, L., Savarese, S.: Social lstm: Human trajectory prediction in crowded spaces. In: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition. (2016)
4. Spinello, L., Luber, M., Arras, K.O.: Tracking people in 3d using a bottom-up top-down detector. 2011 IEEE International Conference on Robotics and Automation (2011) 1304–1310
5. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The kitti dataset. International Journal of Robotics Research (IJRR) (2013)
6. Pandey, G., McBride, J.R., Eustice, R.M.: Ford campus vision and lidar data set. I. J. Robotics Res. **30** (2011) 1543–1552
7. Yan, Z., Duckett, T., Bellotto, N.: Online learning for human classification in 3d lidar-based tracking. In: In Proceedings of the 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Vancouver, Canada (September 2017)
8. Huang, X., Wang, P., Cheng, X., Zhou, D., Geng, Q., Yang, R.: The apolloscape open dataset for autonomous driving and its application (2018)

9. Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., Vasudevan, V., Han, W., Ngiam, J., Zhao, H., Timofeev, A., Ettinger, S., Krivokon, M., Gao, A., Joshi, A., Zhao, S., Cheng, S., Zhang, Y., Shlens, J., Chen, Z., Anguelov, D.: Scalability in perception for autonomous driving: Waymo open dataset (2019)
10. Geyer, J., Kassahun, Y., Mahmudi, M., Ricou, X., Durgesh, R., Chung, A.S., Hauswald, L., Pham, V.H., Mühlegg, M., Dorn, S., Fernandez, T., Jänicke, M., Mirashi, S., Savani, C., Sturm, M., Vorobiov, O., Oelker, M., Garreis, S., Schuberth, P.: A2d2: Audi autonomous driving dataset (2020)
11. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving (2019)
12. Kesten, R., Usman, M., Houston, J., Pandya, T., Nadhamuni, K., Ferreira, A., Yuan, M., Low, B., Jain, A., Ondruska, P., et al.: Lyft level 5 av dataset 2019. urlhttps://level5. lyft. com/dataset (2019)
13. Achlioptas, P., Diamanti, O., Mitliagkas, I., Guibas, L.: Learning representations and generative models for 3D point clouds. In Dy, J., Krause, A., eds.: Proceedings of the 35th International Conference on Machine Learning. Volume 80 of Proceedings of Machine Learning Research., Stockholmsmässan, Stockholm Sweden, PMLR (10–15 Jul 2018) 40–49
14. Yang, Y., Feng, C., Shen, Y., Tian, D.: Foldingnet: Point cloud auto-encoder via deep grid deformation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (June 2018)
15. Teichman, A., Levinson, J., Thrun, S.: Towards 3d object recognition via classification of arbitrary object tracks. In: International Conference on Robotics and Automation. (2011)
16. Zhou, Y., Tuzel, O.: Voxelnet: End-to-end learning for point cloud based 3d object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (June 2018)
17. Carlevaris-Bianco, N., Ushani, A.K., Eustice, R.M.: University of Michigan North Campus long-term vision and lidar dataset. International Journal of Robotics Research **35**(9) (2015) 1023–1035
18. Fan, H., Yang, Y.: Pointrnn: Point recurrent neural network for moving point cloud processing (2019)